Live Class 4

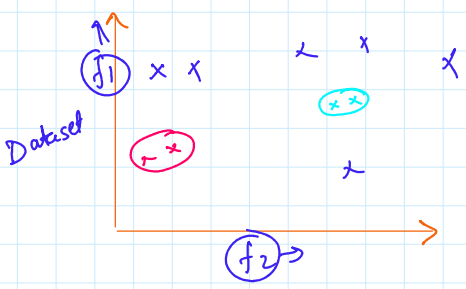
12 February 2024 20:44

Summary → Unsupervised Learning [No y label]
→ Clustering (general method)
→ K-means → initialise / Assign the points / Update
→ Metrics (Tools) → D-I / I.C / Intra / WCSS / S.score
→ K Means ++ (far away)

Hierarchical Clustering → Agglomerative (✓) (Bottom Up)
→ Divisive (opposite of Agglomerative) (Top Down)

What is Agglomerative Clustering?

- The word agglomerative suggests combining things
- It is a **bottom-up** approach
- Agglomerative clustering starts with the assumption that every datapoint is a cluster
- Then, it groups the clusters which are closed to each other until there is only a single cluster left



- Algo

t = 0 → n pts = n clusters

t = t1 (arbitary) → combine some pts → n1 (< n)

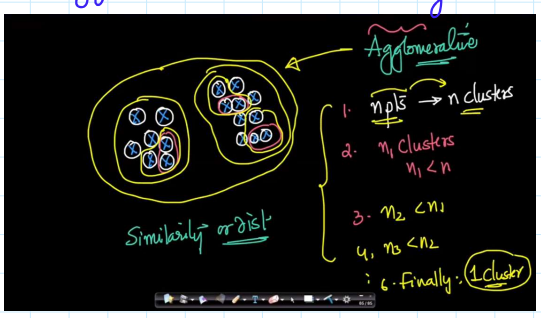
n2 (< n1) → merge / combine more points

steps

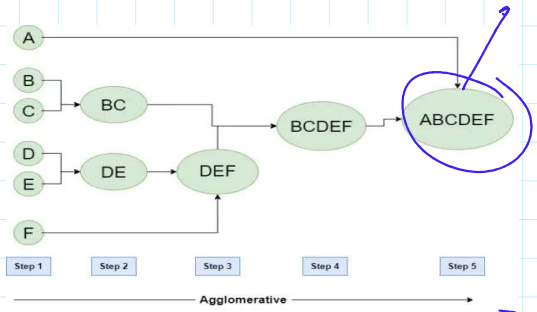
(1 big cluster)

one big cluster

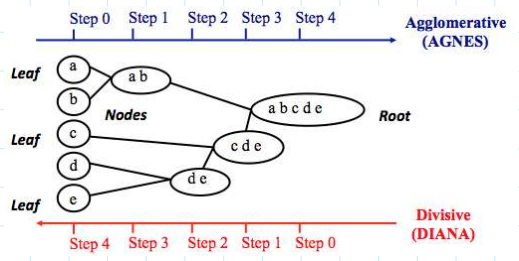
Agglomerative Clustering

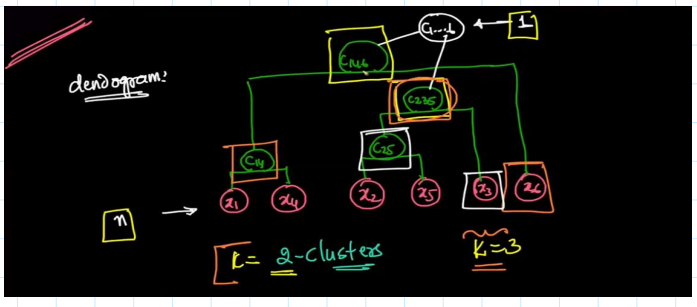


one Big cluster

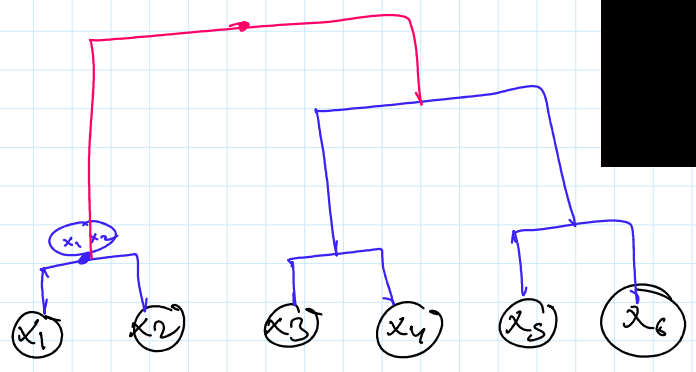


I don't have to specify (k)





centroid



① distance every point x_1 → $x_2 x_3 \dots x_6$

x_2 → every other pts

** $^NC_2$

Divisive Clustering →

What is Divisive Clustering?

- It is an complete opposite of agglomerative approach
- It is a **top-down** approach
- It starts with one big clusters that conatins all the datapoints.
- It then divides the points into different clusters till each data point is a cluster itself

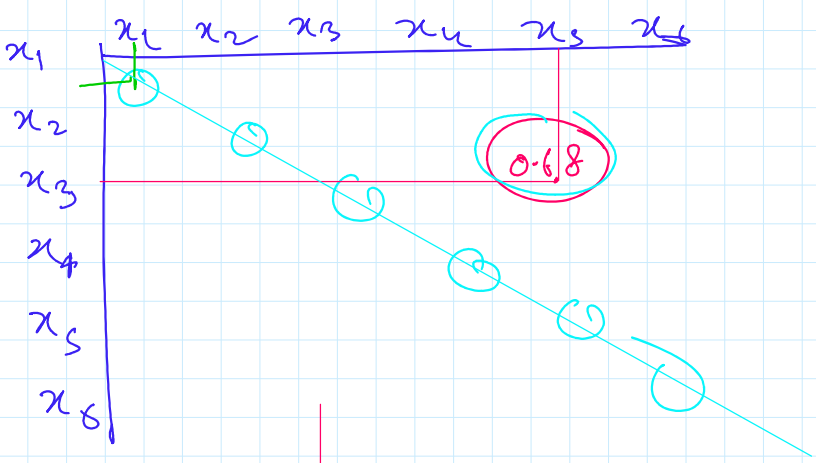
Proximity Matrix → matrix of distance (similarity)

input = A list of points = $[x_1\ x_2\ x_3\ x_4\ x_5]$

Output = $n \times n$ Matrix where each (i,j) represents distance b/w i^{th} and j^{th} clusters (or data point)

small example

Data = $x_1\ x_2\ x_3\ x_4\ x_5\ x_6$ (Pairwise Distance)

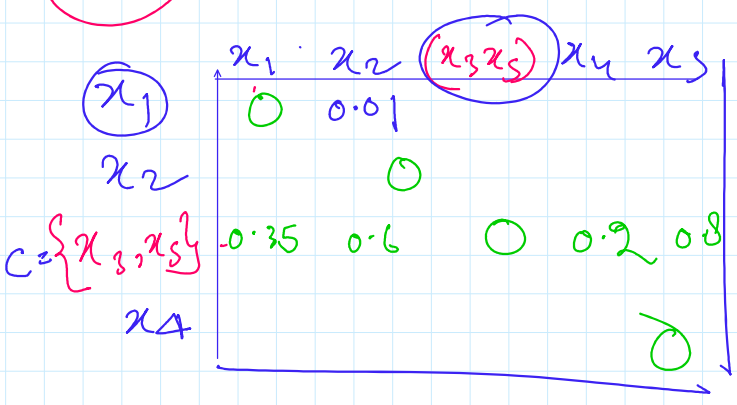


Dist (x_1, x_2) = Dist (x_2, x_1)

* Dist $(x_3, x_5) = 0.68$

merge x_3 and x_5

MIN



$euc = x_2, x_1 = \sqrt{(x_2-x_1)^2 + (y_2-y_1)^2}$

Distance b/w Cluster $(x_3 x_5)$ and x_2

① MIN

② MAX

Group Average

Others

x_2

$x_3\ x_5$

x_2

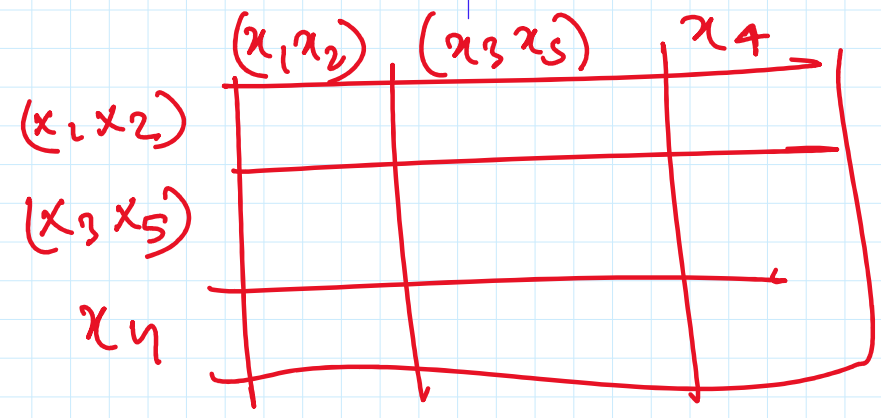
x_3, x_5 x_2

$\min\left[d(x_3,x_2),\ d(x_5,x_2)\right]$

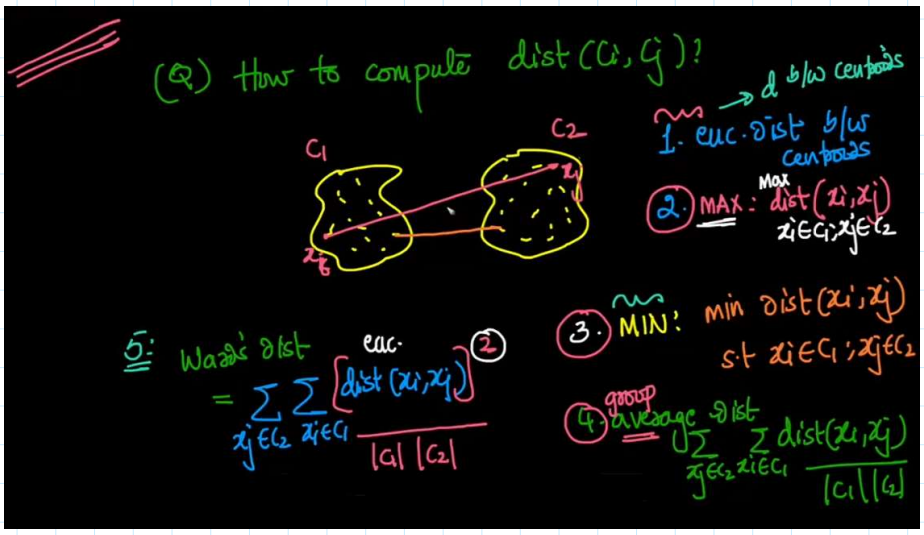
x_3, x_5 x_2

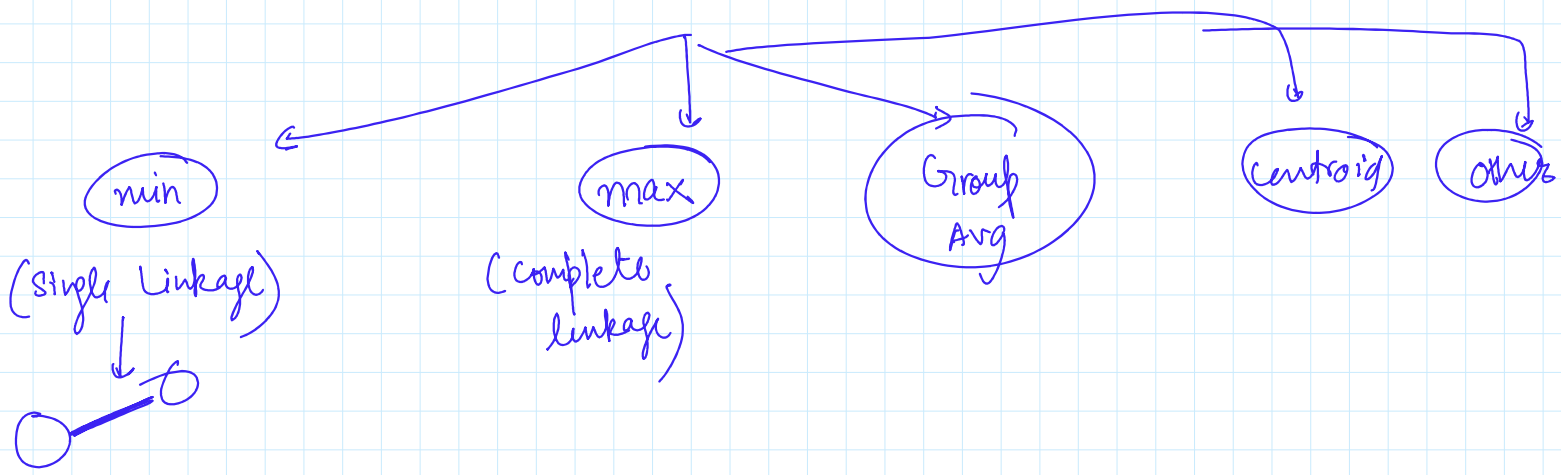
$\max\left[d(x_3,x_2),\ d(x_5,x_2)\right]$

$$\frac{d(x_3,x_2) + d(x_5,x_2)}{2}$$



Distance Metrics



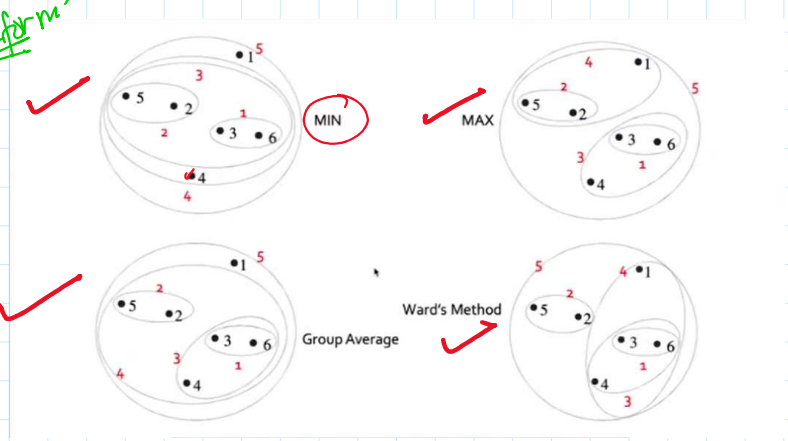


Break → 10:18 pm → 6 min break

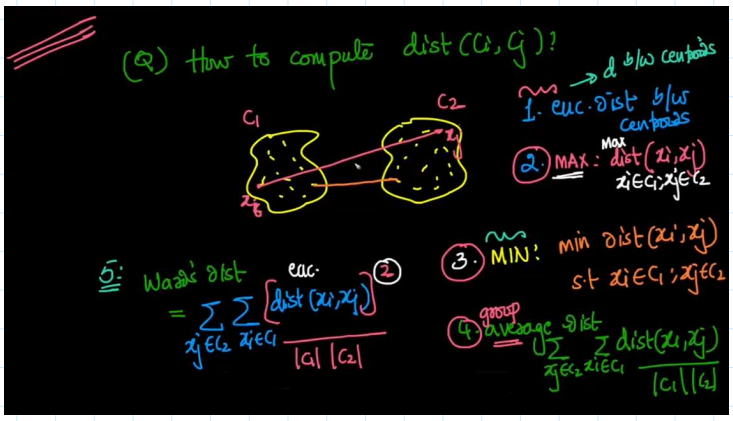
Q How will the distance affect?

Real → (It depends on the data // Domain Specific)

No Inform^n



Ward Distance

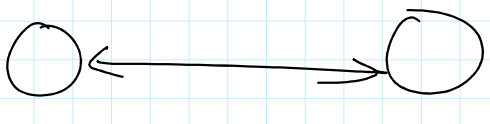


W.D → $\sum_{x_i \in C_1} \sum_{x_j \in C_2} \frac{[dist(x_i, x_j)]^2}{|C_1|\,|C_2|}$

eudedian

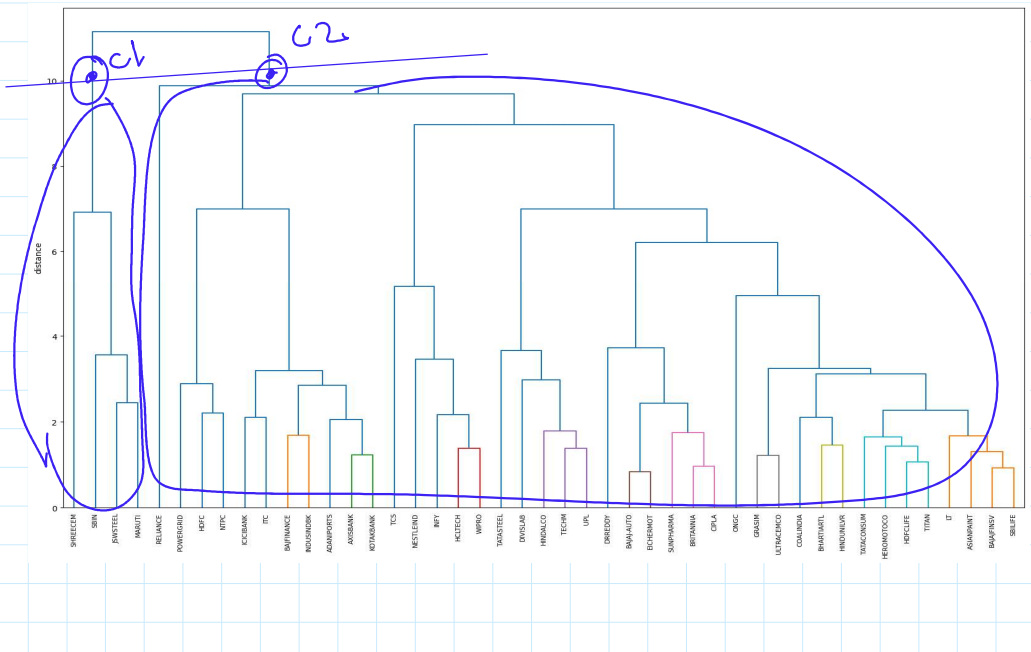
$\text{euclidian}^2 = \left(\sqrt{(x_2-x_1)^2 + (y_2-y_1)^2}\right)^2 = d$

⇒ ward Distance $= \frac{(x_2-x_1)^2 + (y_2-y_1)^2}{|C_1|\,|C_2|}$

Dissimilarity 

Financial Terms

1. Market Cap (Market Capitalization): This is the total market value of a company's outstanding shares of stock. It is calculated by multiplying the current market price of one share by the total number of outstanding shares. It's an indication of the company's size and has implications for how volatile the stock price might be and how broad the ownership might be.
2. Regular Market Volume: This refers to the number of shares of a stock that are traded during regular trading hours on a given day. It is an indicator of the stock's liquidity and the level of investor interest in the company.
3. Quarterly Earnings Growth: This metric shows the percentage increase or decrease in a company's earnings per share (EPS) from one quarter to the next. It is an important indicator of a company's profitability and its potential to increase revenue and manage costs over time.
4. Book Value: The book value of a company is calculated from the balance sheet, and it's the difference between a company's total assets and total liabilities. It represents the net asset value of the company according to accounting principles.
5. Total Revenue: This is the total amount of income generated by the sale of goods or services related to the company's primary operations. It is often referred to as the "top line" because it is the first number you see on a company's income statement.
6. Return on Assets (ROA): This is a profitability ratio that measures how effectively a company can earn a return on its investment in assets. It is calculated by dividing the company's net income by its total assets.
7. Profit Margins: This financial metric measures the percentage of revenue that exceeds the costs associated with a company's products or services. There are various types of profit margins, including gross, operating, and net profit margins, each providing different levels of cost and profit insight.
8. Earnings Growth: This represents the annual compounded growth rate of earnings from investments over time. It indicates the long-term earning potential and financial stability of a company.



Determining No. of Clusters with Dendrogram

If you want to create flat clusters we can analyze the above dendrogram to determine no. of clusters. We first assume that the horizontal lines are extended on both sides, and as such, they would also cross the vertical lines. Now we have to identify the tallest vertical line that does not have any horizontal line crossing through it.

In the above dendrogram graph, such a vertical line is the blue line. We now draw a horizontal line across this vertical line as shown below. This horizontal line cuts the vertical line at two places, and this means the optimal number of clusters is 2.

Another way is to visually see which vertical line is showing the biggest jump. Since the vertical line denotes the distance or similarity between the two clusters, the big jump signifies the two clusters are not very similar. Again draw the horizontal line through this vertical line and the number of cuts it makes is optimal no. of clusters. Again in our example, it is the blue line, and the horizontal line cuts at two places so no. of clusters is 2.

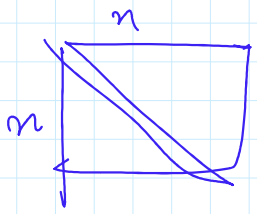
(It should be noted however that these methods do not always guarantee the optimal number of clusters, it is just a guideline)

Limitations of Agglomerative Clustering

49 stocks → 49 x 49

1M data → 1M x 1M → computationally intensive

Space Complexity → $O(N^2)$



Time Complexity → $O(N^3)$

$(10^7)^3$ → hours